SciComp 2007

Harnessing Massive Parallelism in the Era of Parallelism for the Masses

David Klepacki
IBM T.J. Watson Research Center

Questions

- **How will the multicore revolution impact supercomputing software?**
- **How will today's HPC community impact multicore software and applications?**
- **Will we take advantage of ever increasing numbers of cores per chip, or will the multicore revolution stall?**

Outline

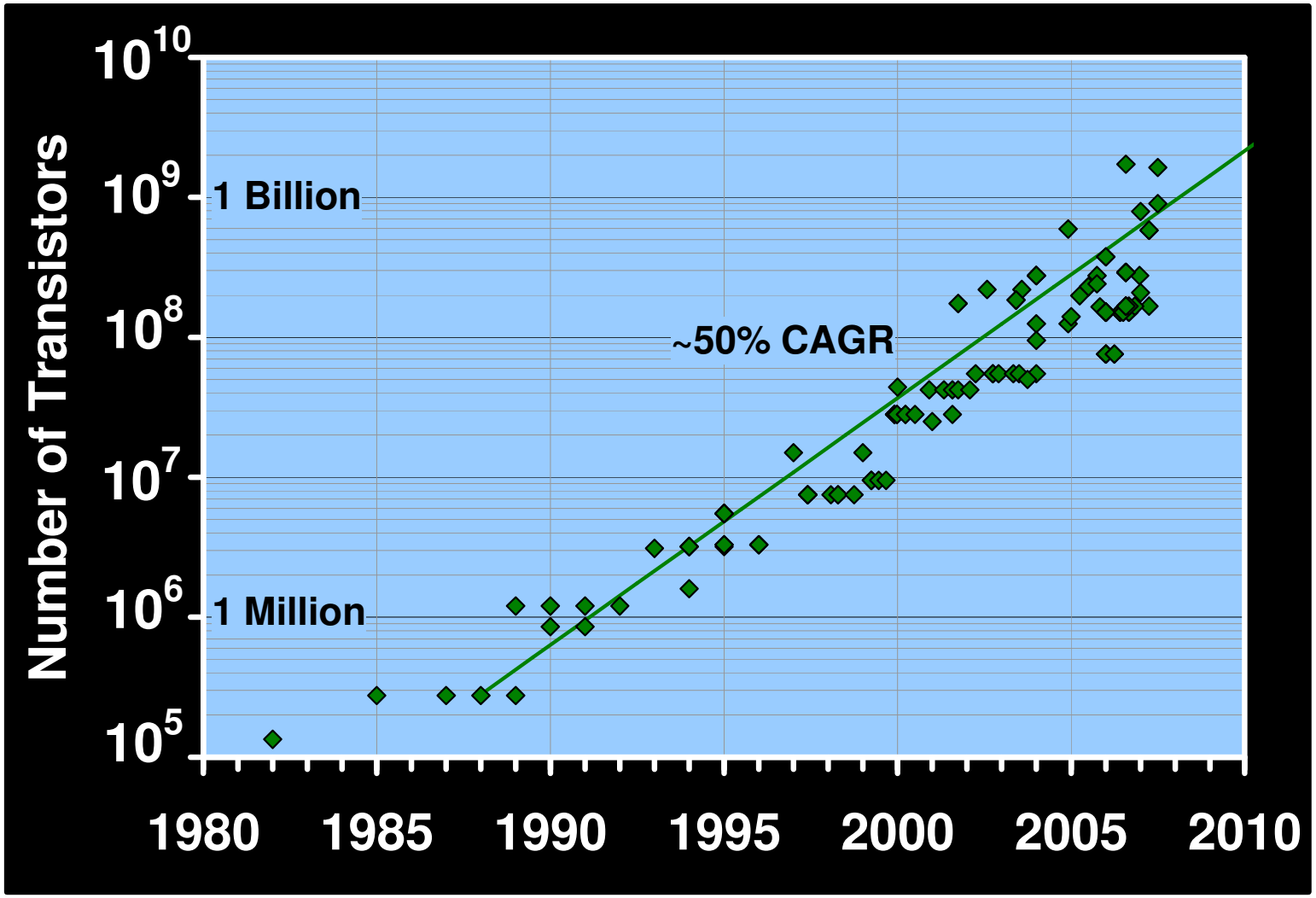
- **System Architecture Directions**
- **Challenges for Petaflop supercomputing**
- **Supercomputing Software Directions**
  - Programming languages
  - Operating systems
  - Tools
  - Education

A New Era in Systems Design

- **Highly parallel systems built with multiple small processors are addressing a growing fraction of the server market**
  - Generally, they are characterized by weak single-thread performance, good chip-level throughput performance and excellent power-performance
- **Parallelism levels that once were only within the high performance computing (HPC) domain will become mainstream and will be exploited at all levels of the software stack**

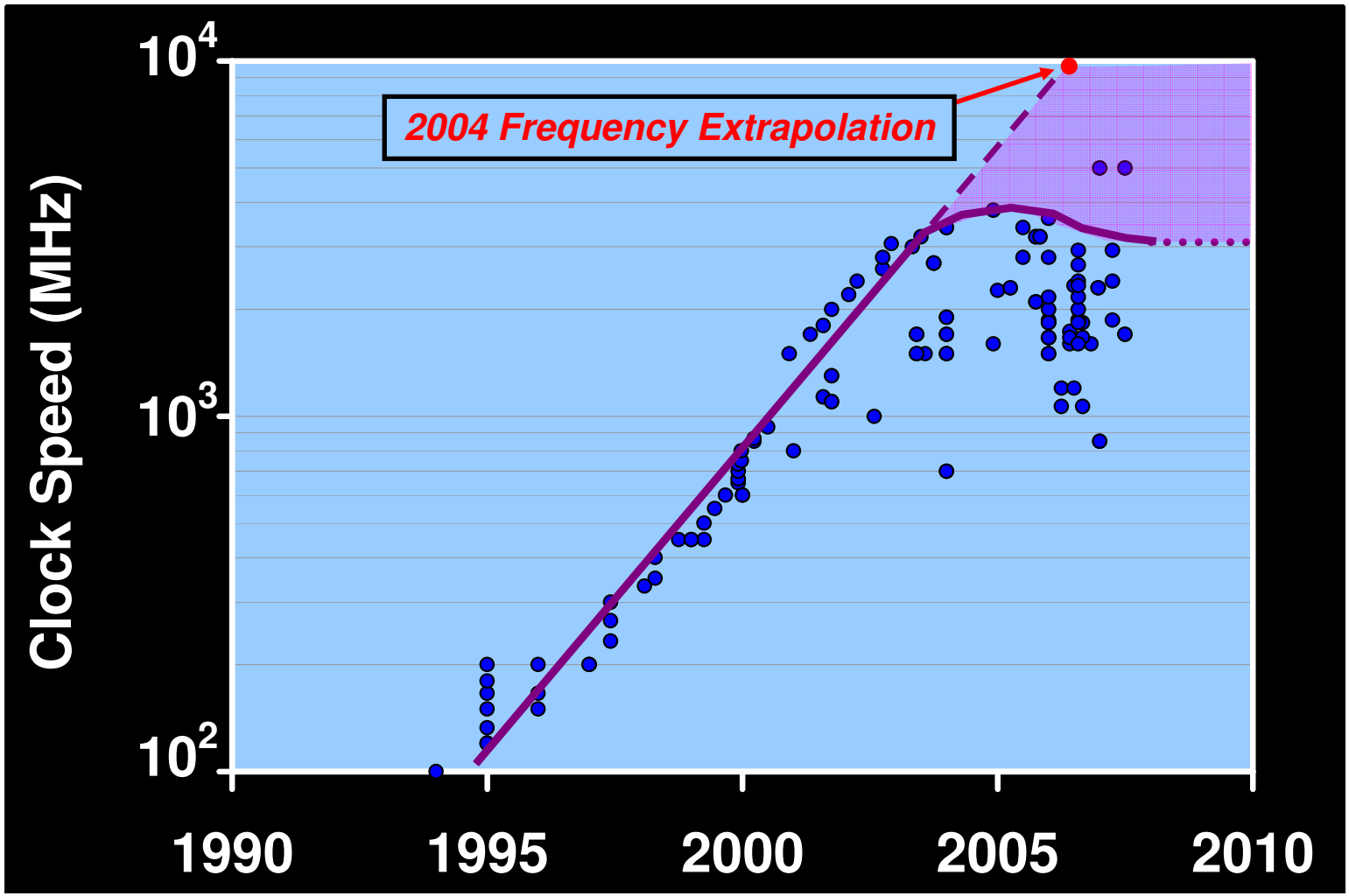
Microprocessor Transistor Trend

Lithography will continue to deliver density scaling



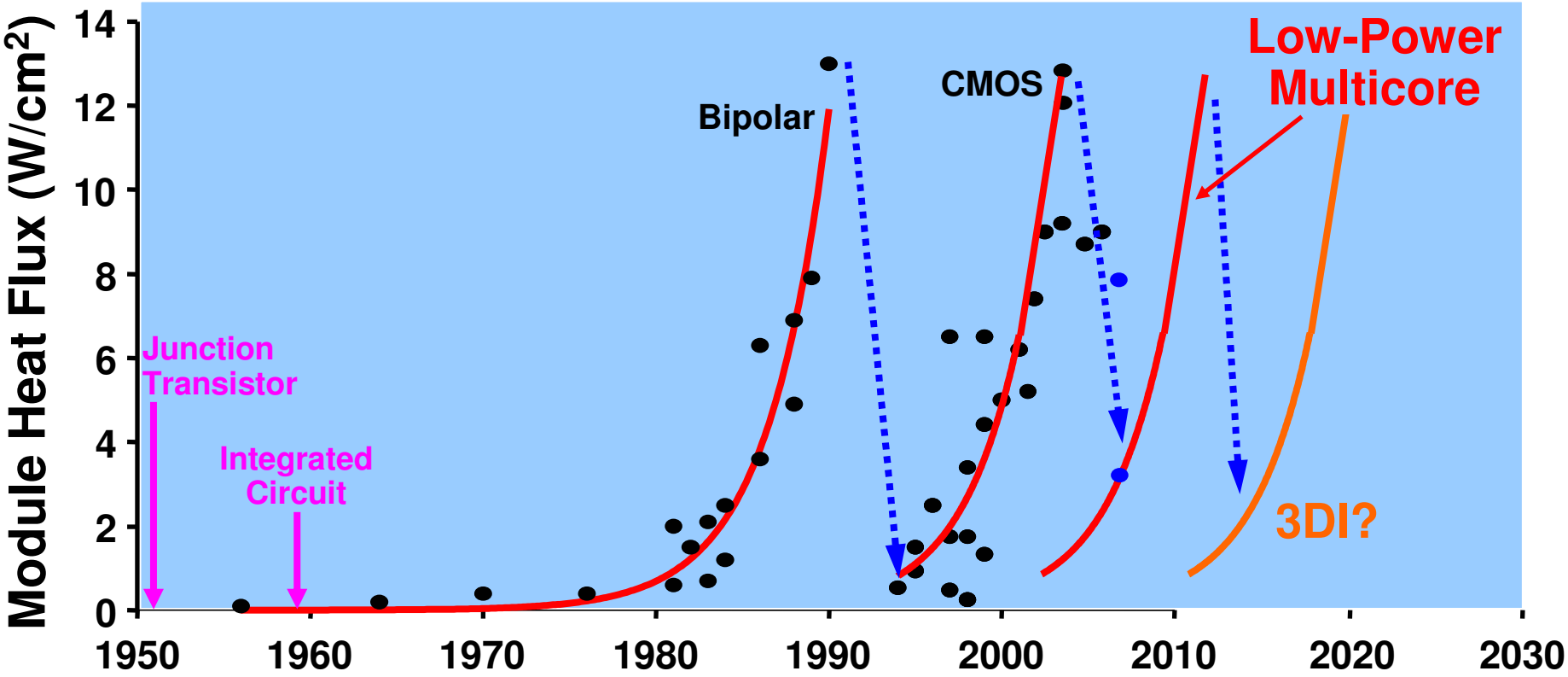
Microprocessor Clock Speed Trends

Managing power dissipation is limiting clock speed increases



Chip Power Density

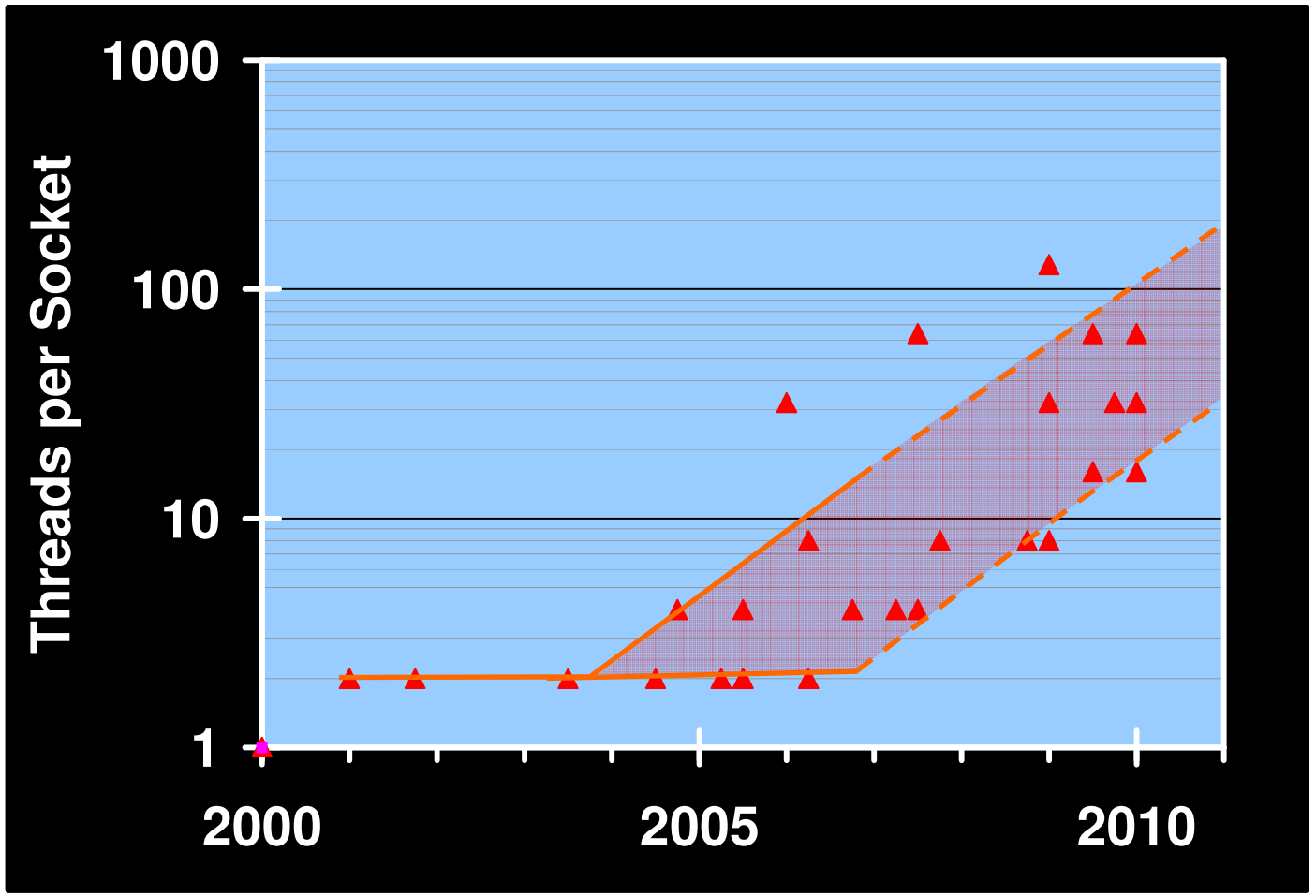
Power is constraining frequency scale-up leading to the emergence of lower power multicore chips



Low power multicores replacing single high-power cores

Server Microprocessors Thread Growth

We are entering a new era of massively multi-threaded computing



Hardware trends that address the power problem

- **Observation**
  - **Although frequency scaling is "dead", Moore's Law is still alive: transistor density continues to increase exponentially**
- **Trend #1: Multi-core processor chips**
  - **Maintain (or even reduce) frequency while replicating cores**
- **Trend #2: Accelerators**
  - **Previously, processors would "catch" up with accelerator function in the next generation**
    - **Accelerator design expense not amortized well**
  - **New accelerator designs will maintain their speed advantage**
  - **And will continue an enormous power advantage for target workloads**

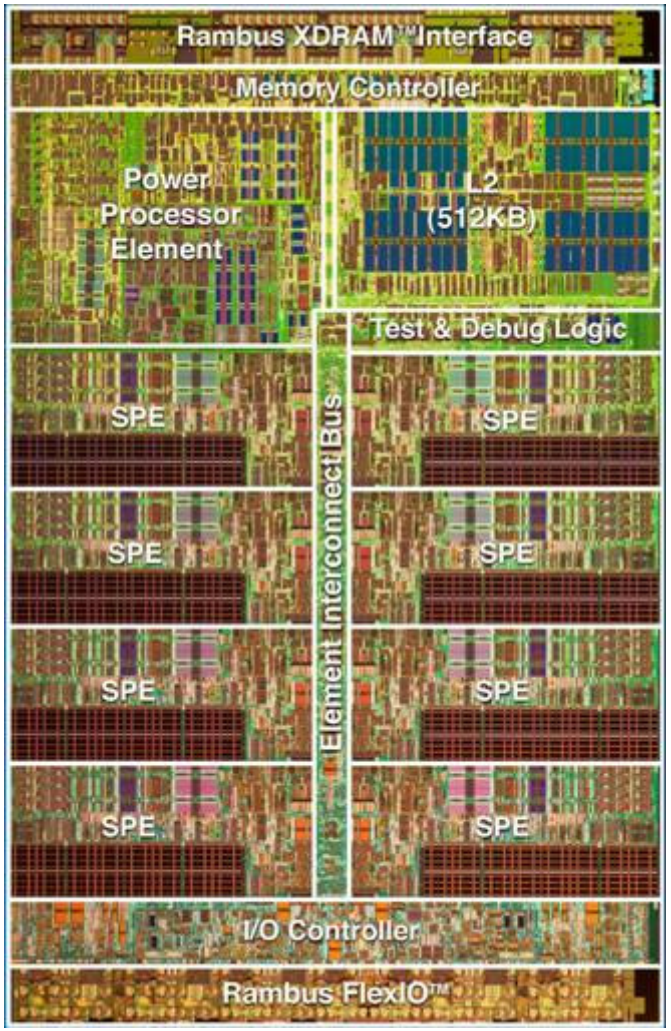
Blue Gene/L, an example of addressing power in a massive scale-out system

- **128K compute processors**
  - **367 Teraflop Peak**
  - **280 Teraflop Linpack**
- **3D torus interconnect**
- **Collective and barrier networks**
- **Power: 1.5 MW**
  - **1.5 MW**
  - **>0.25 Gigaflop/W!**
- **Area**
  - **64 compute racks**
  - **2500 ft^2 (232m^2)**

BG/L at LLNL, #1 on the Top500 list

Cell BE highlights - 3.2 GHz

- 241M transistors
- 235mm^2
- 9 cores, 10 threads
- >200 GFlops (SP)
- >20 GFlops (DP)
- Up to 25 GB/s memory B/W
- Up to 75 GB/s I/O B/W
- >300 GB/s EIB
- Throughput per rack 16Tflops
- Sustained throughput on FFTs 50 Tflops (25%)



Software trends that complement hardware trends

- **Trend #1: Applications are becoming more scalable**
  - **Many new applications are inherently scalable:**
    - Streaming apps
    - Sensor apps
    - Search-based apps
  - **Examples**
    - Digital Video Surveillance
    - Web servers
    - Rich Media mining
  - **Many old applications, middleware, and operating systems are being modified to be more scalable**
    - Some will prove very difficult to modify
    - In fact, not every app should be parallelized

Software trends that complement hardware trends

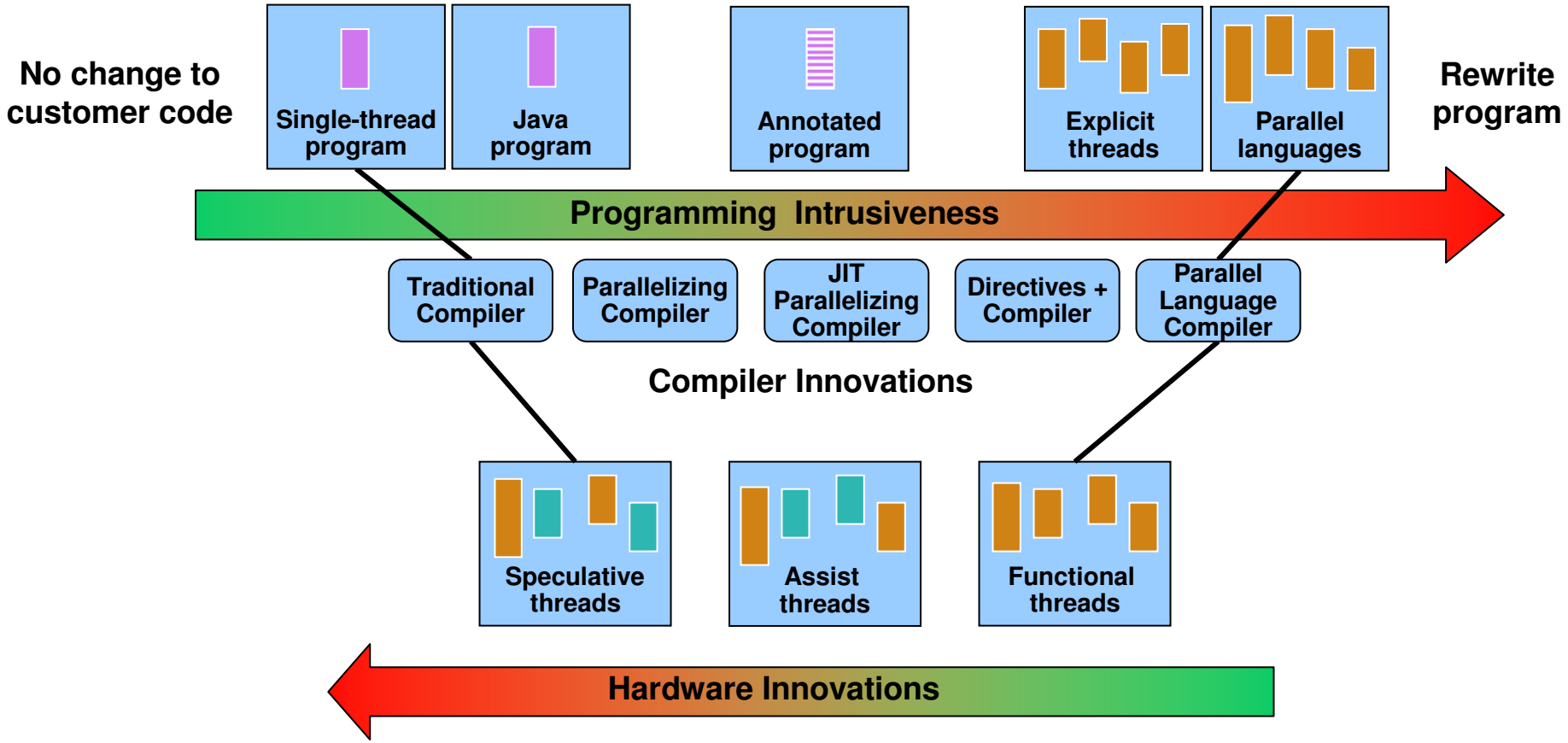
- **Trend #2: Partitioned global address spaces**
  - **Global addressing key to many programming models**
    - Simpler to program
  - **For scaling, still crucial to know/control data location**
  - Although, don't forgo the ability to pass messages

Software trends that complement hardware trends

- **Trend #3: Ecosystem for parallel apps is improving**
  - **Compilers**
    - Improvements are incremental
  - **Languages**
  - **Performance analysis tools**
  - **Debuggers?**
    - For highly parallel, need something better than printf!
  - **IDEs**
    - Although many HPC applications experts aren't interested

Different approaches to exploit multicore chips

Systems built around multicore processor chips are driving the development of new techniques for automatic exploitation by applications



Sustained Petaflop machines

- **Several systems with over 100 Teraflops have been installed to date**
- **Peak petaflop systems will appear in 2008**
- **US National Science Foundation (NSF) "Track 1" acquisition: sustained Petaflop system by 2011**
  - Submitted bids have peaks far exceeding 1 Petaflop

The Billion Dollar Race to a PetaFlop

- **Japan**
  - Effort to regain number one position
  - PetaFlop in 2008
  - 3-4 PetaFlops in 2010 (official target)
    - 10 PF (unofficial target)
  - 100's of PF in 2015-2020

京 速 計算機

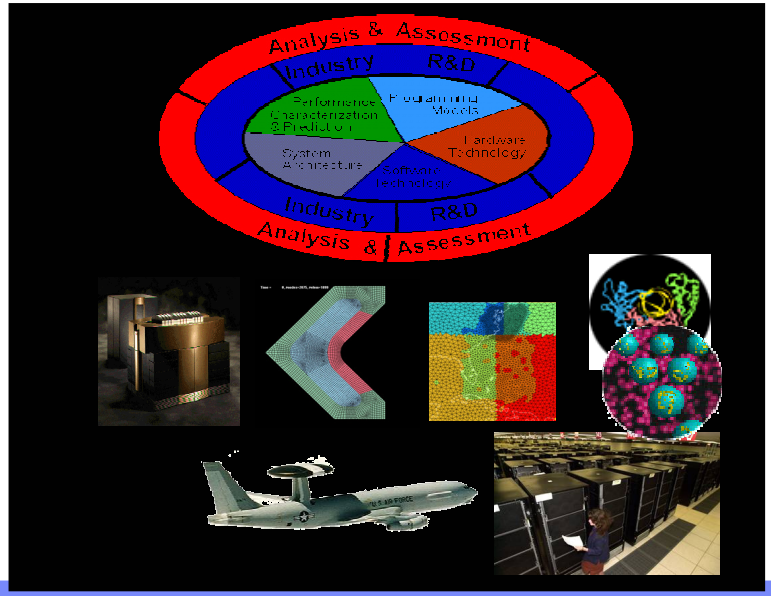
Kei *Soku* *Keisanki*

10^{16} speed computer

Formal Title:

"Development of most advanced high-performance general-purpose supercomputer"

- **US**
  - DARPA: HPCS program - Develop: Multi - PetaFlops highly productive system in 2010-2011
    - Cray, IBM, Sun
  - NSF: To announce program to procure >PF sustained system in 2010
  - Large DOE labs looking for something before 2010



IBM to Build World's First Cell Broadband Engine™ Based Supercomputer

Revolutionary Hybrid Supercomputer at Los Alamos National Laboratory Will Harness Cell Game Chips and AMD Opteron™ Technology

x86 Linux Master Cluster
AMD Opteron™ System x3755

Cell BE Accelerator Linux Cluster
8000+ Blades

- **Goal 1 PetaFlop Double Precision Floating Point *Sustained***
  - *1.6 PetaFlop Peak DP Floating point (3.2 SP)*
  - *360 server racks that take up around 12,000 square feet--about three basketball courts.*
  - *Hybrid of Opteron X64 AMD processors (System x3755 servers) and Cell BE Blade Servers connected via high speed network*
  - *Modular Approach means the Master Cluster could be made up of Any Type System – including Power, Intel*

Los Alamos NATIONAL LABORATORY

Programming models for the Petascale era

- **Today, most highly scalable codes use message passing**
  - Typically some variant of MPI + OpenMP
- **Other programming models exist**
  - For example, Charm++, RapidMind, PeakStream (RIP)
- **Increased usage of PGAS languages**
  - Global address space, with explicit locality
  - For example, UPC is ramping up in usage
  - DARPA-funded efforts. For example IBM's X10
- **Users will enable/annotate speculative parallelization**
  - Including Transactional Memory

Blue Gene/L: a preview of Petaflop scaling issues?

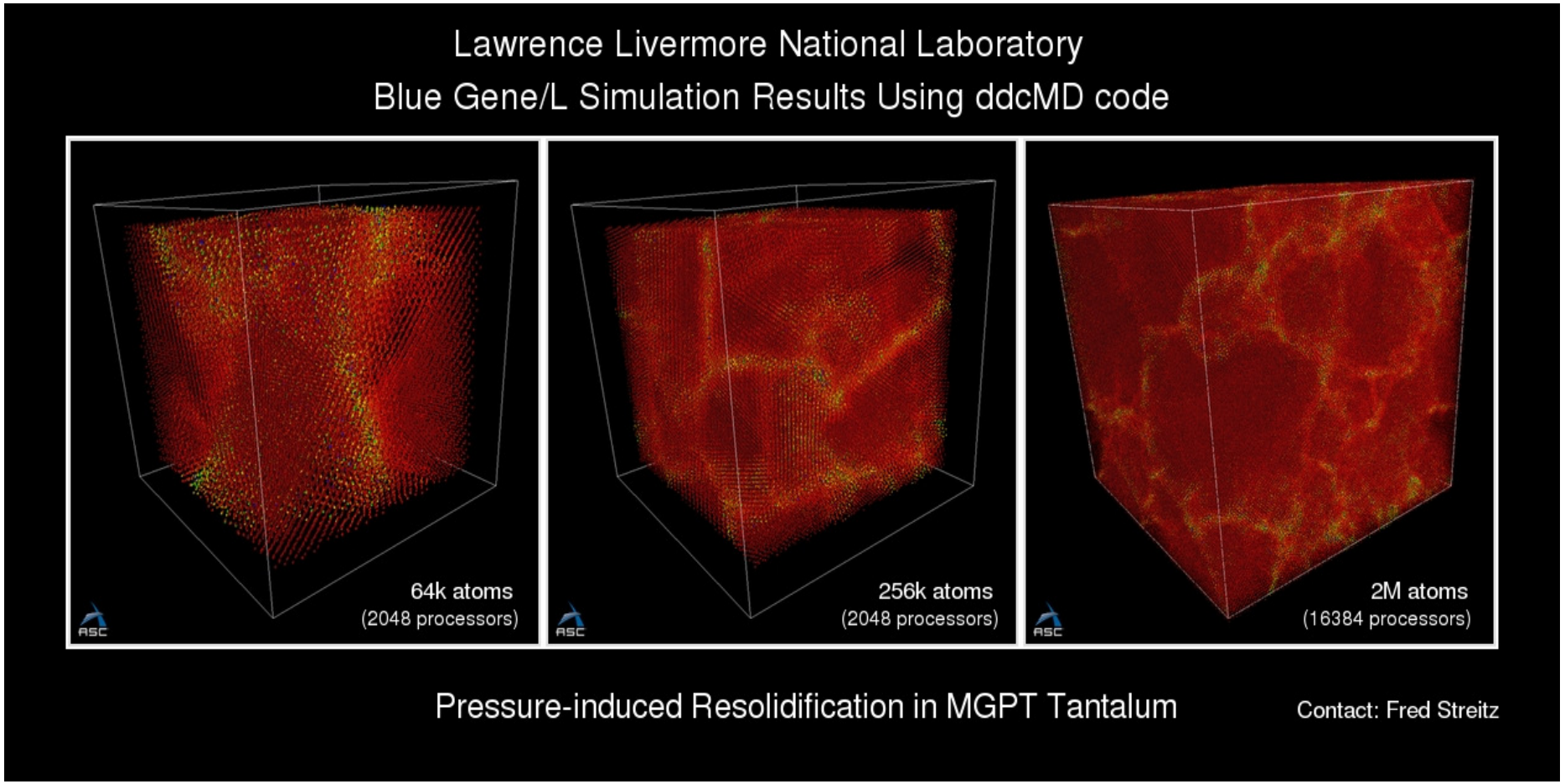
- **Largest Blue Gene/L machine has 128K cores**
  - Petaflop machines based on commodity processors will need a similar number of cores
  - For instance, for 3.3 GHz cores with 2 flop/cycle, would need 150K cores to reach one Petaflop

- **Blue Gene has scaled application performance to up to 128K cores for a number of important scientific apps**

Examples of highly scalable applications running on BG/L and ASC Purple

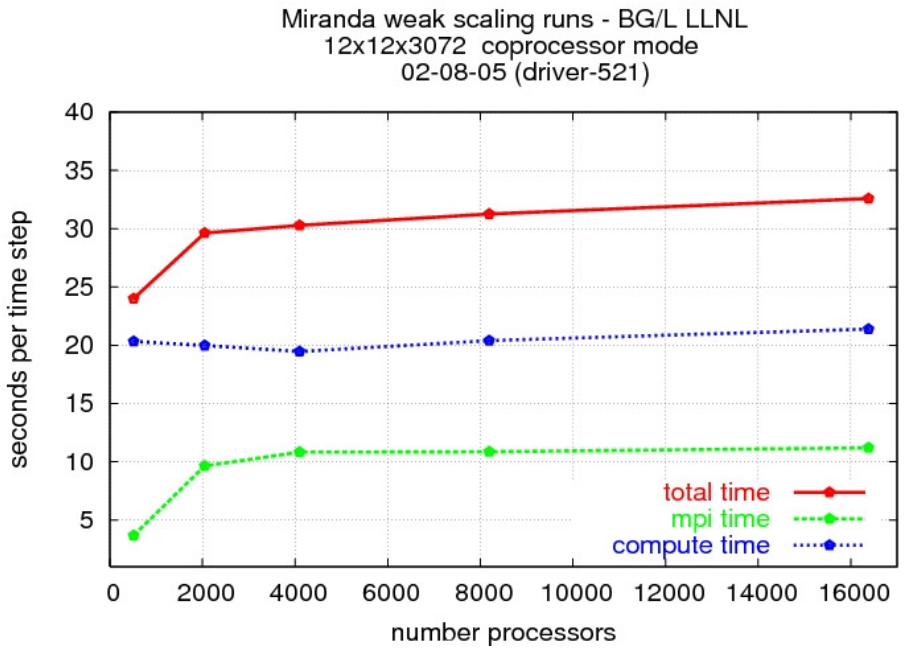
- **High-Energy Physics:**
  - **FDTD Nanophotonics**
  - **GTC (Princeton University)**
  - **QCD kernel (Jurelich, Columbia, IPP)**
  - **SPHOT (LLNL)**
- **Astro Physics**
  - **Flash (U Chicago, ANL)**
- **Atmospheric/Ocean modeling, Weather & Climate**
  - **HOMME (NCAR)**
  - **POP (LANL, ANL)**
- **Material Science**
  - **ParaDis (LLNL)**
- **Molecular Dynamics**
  - **Blue Matter (IBM)**
  - **NAMD (UUIC/NCSA)**
  - **CPMD (IBM)**
  - **SpaSM (LANL)**
  - **LAMMPS (SNL)**
  - **ddcMD (LLNL)**
  - **Qbox (LLNL)**
  - **Grasp (SNL)**
  - **MDCASK (LLNL)**
  - **DL-POLY v3 (CCLRC, Daresbury)**
- **Computational Chemistry**
  - **Shake & Bake (U Buffalo)**
- **Computational Fluid Dynamics**
  - **NEK5000 (ANL)**
  - **AVBP (CERFACS)**
  - **Miranda (LLNL)**
  - **Raptor (LLNL)**
  - **SAGE (LLNL, SAIC)**
- **Reservoir Simulation**
  - **IPARS (U TX Austin)**
- **Libraries & Tools**
  - **ScaLAPACK (ONL, U Tennessee)**
  - **PETSc (ANL)**
  - **Total View (Etnus)**
- **Benchmarks**
  - **GUPS**
  - **HPL-LINPACK**
  - **UMT2K**
  - **Sweep3D**

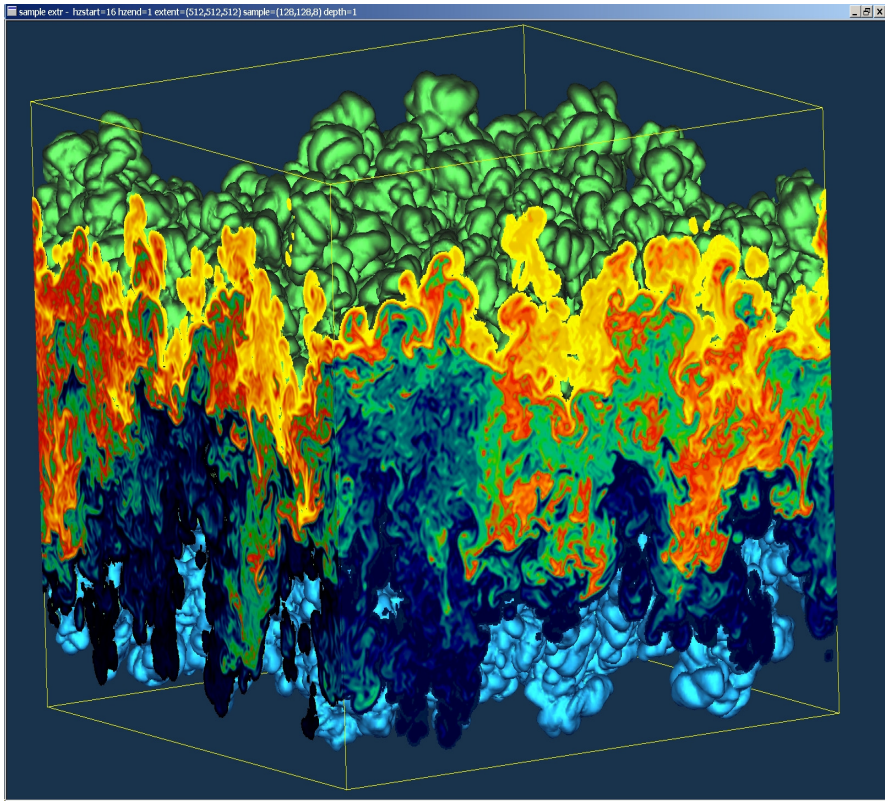
Classical MD – ddcMD



64K and 256K atom simulations on 2K nodes were already an order of magnitude larger than previously attempted!

Instability and Turbulence - Miranda





High order hydrodynamics code for computing fluid instabilities and turbulent mix

Applications now Running on BG/P

Application	SMP/ 1Thread	SMP/ 4Thread	Dual Mode	VNM	Largest Node Count	Comment
Linpack	yes	yes		yes	256	Benchmark
NAS Serial	yes				1	NAS Benchmark
NAS OpenMP		yes			1	NAS Benchmark
NAS Parallel	yes			yes	128	NAS Benchmark
PALLAS	yes				32	Benchmark
STREAM	yes	yes			32	Benchmark
MPPTEST	yes			yes	32	Benchmark
MILC	yes			yes	32	NSF Benchmark, QCD Collaboration
PARATEC	yes			yes	32	NSF Benchmark, LBNL
NAMD	yes			yes	512	NSF Benchmark, U Illinois
HOMME	yes			yes	32	NSF Benchmark, NCAR
UMT2K	yes	yes	Yes	yes	512	ASCI-Purple Benchmark
sPPM	yes	yes	Yes	yes	512	ASCI-Purple Benchmark
SPHOT	yes	yes	Yes	yes	512	ASCI-Purple Benchmark
POP	yes			yes	32	LANL
FLASH	yes			yes	32	U Chicago
RAPTOR	yes			yes	32	LLNL code
CPMD	yes	yes		yes	512	IBM
AMR	yes			yes	1024	TI 08 Benchmark
AVUS	yes		yes		1024	TI 08 Benchmark
CTH	yes		yes	yes	512	TI 08 Benchmark
HYCOM	yes		yes		1024	TI 08 Benchmark
ICEPIC	yes			yes	512	TI 08 Benchmark
LAMMPS	yes			yes	1024	TI 08 Benchmark
OOCORE	yes			yes	512	TI 08 Benchmark
WRF	yes			yes	2048	TI 08 Benchmark
GAMESS	yes			yes	1024	TI 08 Benchmark

Current Snapshot of Programming Models (BG/L)

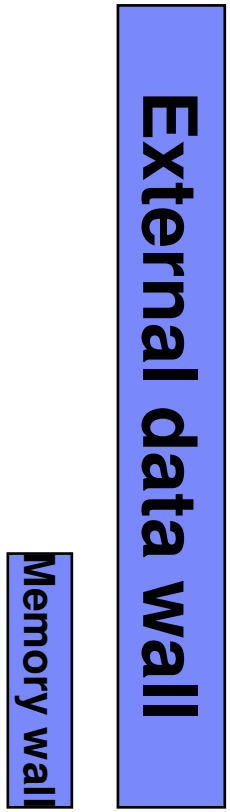
- **MPI is currently the primary programming model**
  - Remember when people worried about more than 1K MPI tasks? How about 128K MPI tasks?
- **Charm++ (for NAMD), ARMCI/GA supported**
- **UPC supported internally**
  - Impressive results for HPC Challenge benchmarks
- **Common lib support: ESSL, MASS/V, glibc, ScaLAPACK, FFTW**

MPP Strategies

- **Leverage low-power cores and System-on-a-Chip (SoC)**
  - **Superior power/performance**
  - **Can pack more performance per rack (and still cool it!)**
- **Keep it simple**
  - **Minimize number/type of chips: SoC**
  - **Simple OS**
- **Address enemies of scalability**
  - **Low latency communication (easier with slower cores ☺)**
  - **Collective communication networks, including barrier sync**
  - **Simple OS (more predictable, less asynchrony)**
  - **Pay close attention to RAS (reliability, availability, serviceability)**

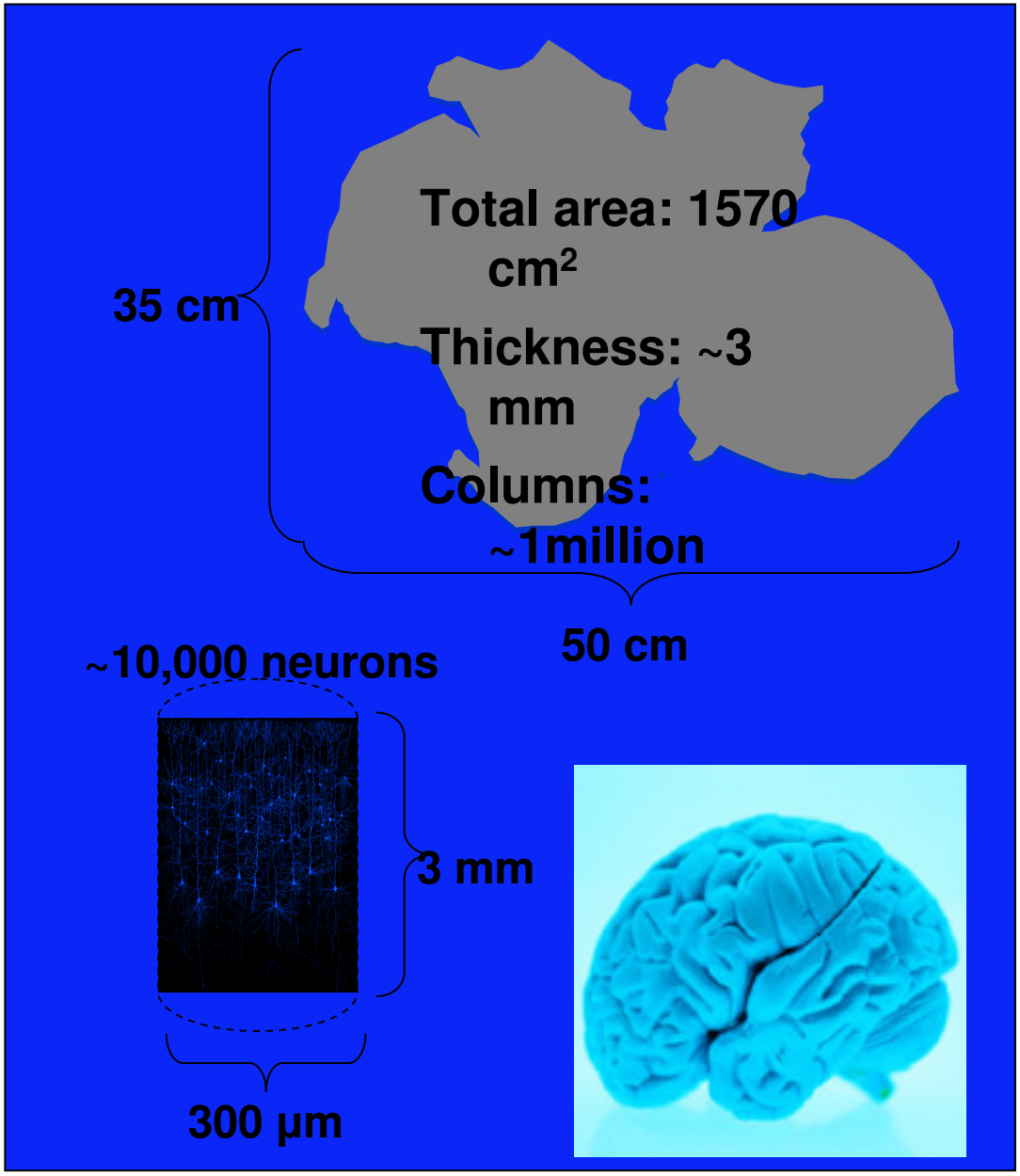
The HPC data challenge

- **Datasets and data models are exploding exponentially in size**
  - Case: Blue Brain: modeling an entire brain
- **The I/O wall is rapidly increasing in size**
  - Disk bandwidth cannot match pace of multicore performance increases
- **Sensor-based data rapidly increasing**
  - Often write-once, read-rarely
  - Case: SKA (Square Kilometer Array)



Blue Brain

- **Joint research collaboration between IBM & EPFL to simulate the** neocortical column
- **Our understanding of the brain is limited by insufficient information and complexity**
  - Overcome limitations of neuroscientific experimentation
- **Inform experimental design and theory**
- **Enable scientific discovery for understanding brain function and diseases**
- **Entire human brain by 2015!**

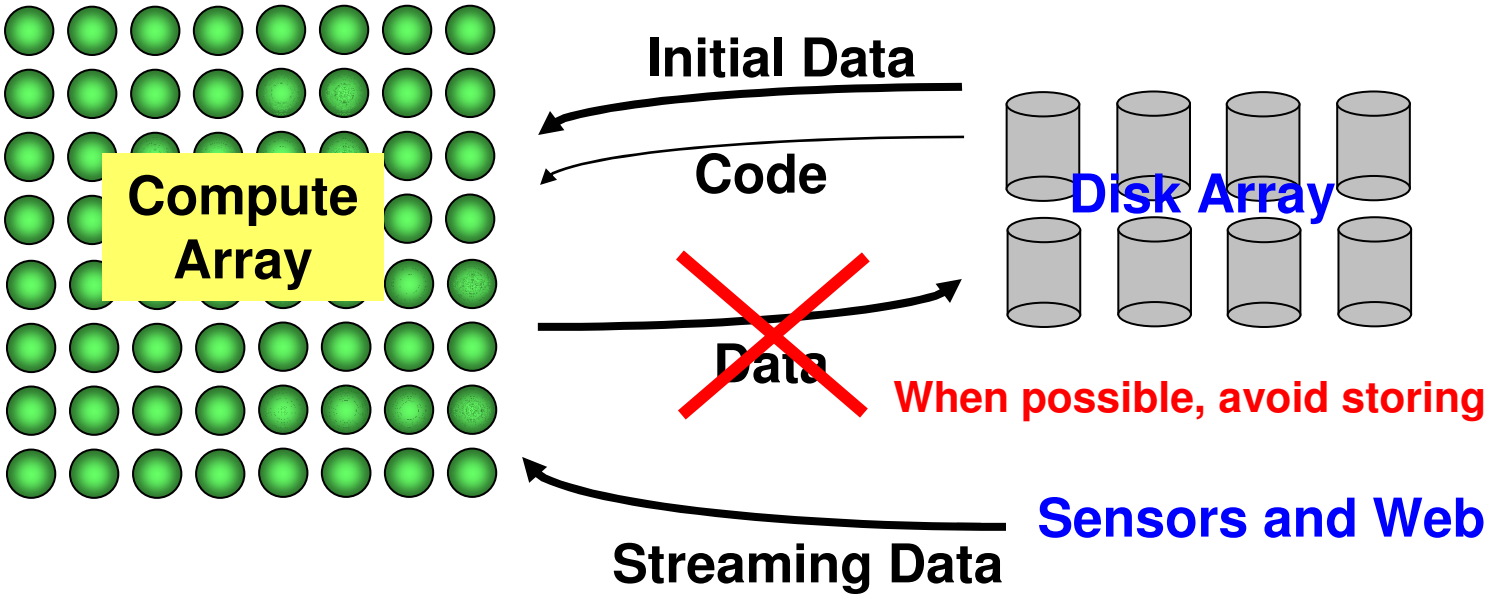


Attacking the HPC data challenge

- **Parallel disk arrays for bandwidth**
  - Expensive and wasteful of disk capacity
- **Disk caching**
  - Not effective for many large HPC applications
- **Compression**
- **For sensor data, filter and analyze on-the-fly**
  - When possible, don't store
- **New "Storage Class Memory" technology**
  - Will Phase Change Memory save the day?

Attacking the HPC data challenge: new directions

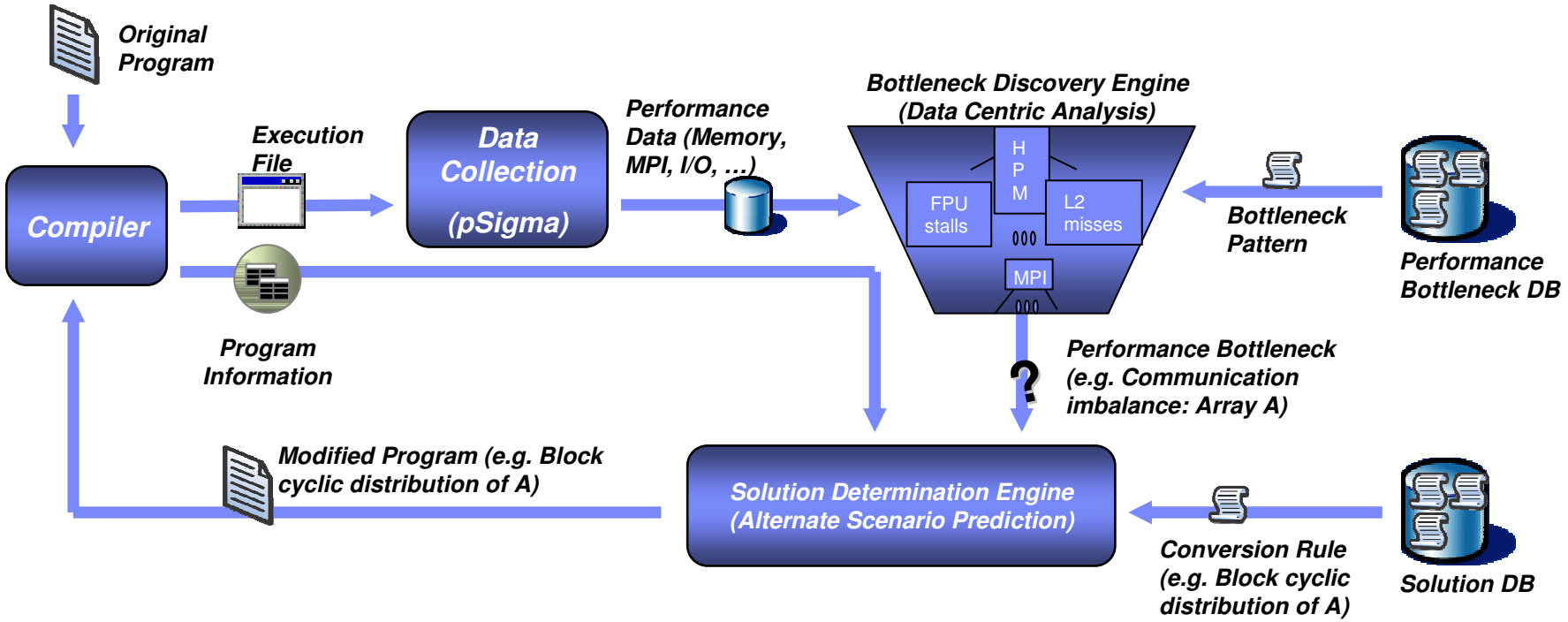
- **Leave the data "in memory" whenever possible**
- **Bring the applications to the data**
- **Visualization/rendering in situ**
- **Issues: App staging, MTBF, memory hogging, …**



DARPA (PERCS) HPCS Toolkit

- **HPCS Toolkit Framework – extension of HPC Toolkit**
  - **Integrated HPM, MPI, SHMEM, and OpenMP tools**
  - **Support for Fortran/95 and C/C++**
- **Technologies to optimize productivity**
  - **Non-intrusive instrumentation and performance data collection**
  - **Bottleneck Discovery Engine (data centric analysis)**
  - **Solution Determination Engine (alternate scenario prediction)**
  - **Solution Implementation Engine (source code patching of solutions)**
- **Enhancements and New Tools**
  - **Front-End Visual Performance Navigator (extension of Xprofiler)**
  - **Memory profiling and analysis (SiGMA)**
  - **Cache visualization**
  - **I/O analysis (extension of MIO)**
  - **Scriptable command line interface or interactive visual framework**

Data and Control Flow of HPCS Toolkit



HPCS Toolkit provides Autonomic Application Performance Capability.

- **Intelligent automation of performance evaluation and decision system**
- **Interactive capability with graphical/visual interface always available, but always optional**

Education

- **Too few software engineers understand how to take advantage of parallelism, particularly data parallelism**
  - Continuation of the multicore revolution is at stake
  - And supercomputers are now dependent upon multicore
- **How to change this?**
  - Update skills through internal and industry courses
  - Introduce parallelism and concurrency early in undergraduate programs
    - But how early (First course? Junior year?)
- **Producing efficient parallel codes can be a challenge**
  - But an exciting and inspiring one!

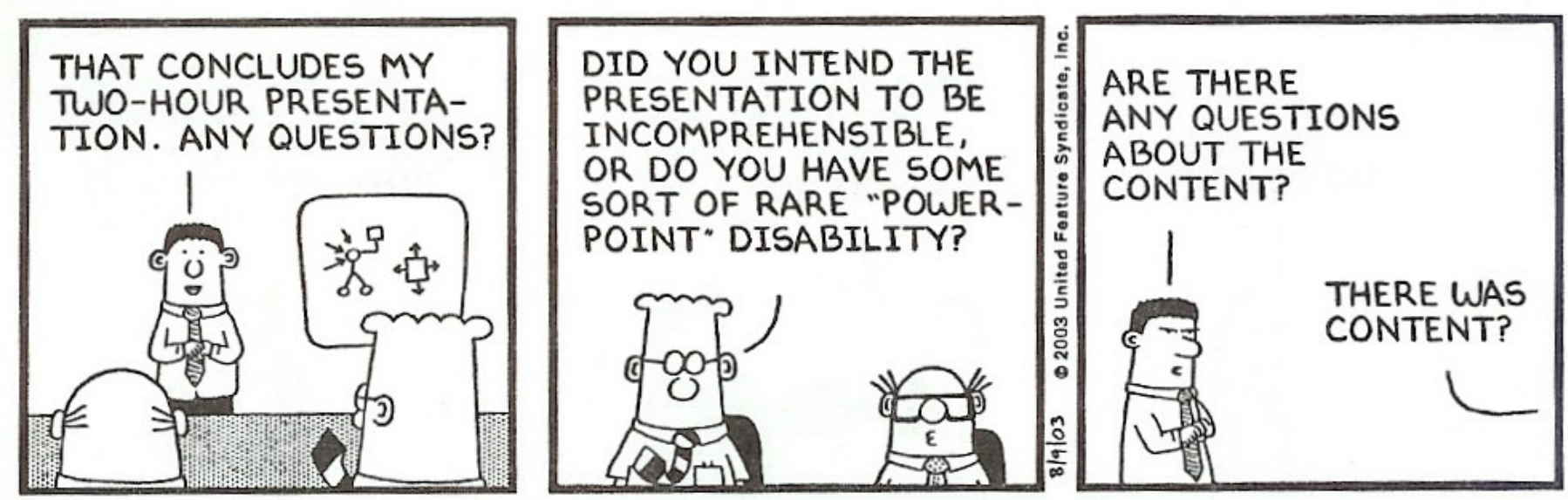
Concluding thoughts

- **How will the multicore revolution impact supercomputing software?**
  - Tools will improve dramatically
  - Parallel language proliferation and experimentation
- **How will today's HPC community impact multicore software and applications?**
  - A wealth of experience to apply. Let's jump in with both feet.
- **Will we take advantage of ever increasing numbers of cores per chip, or will the multicore revolution stall?**
  - Somewhere in between ☺. A caution: continued supercomputing gains likely depend on steady gains in commodity multicore chip performance

Ending on a thoughtful note

- **In the end, it's not about the technology; it's what you do with it that counts**

Questions?


THAT CONCLUDES MY TWO-HOUR PRESENTATION. ANY QUESTIONS?
DID YOU INTEND THE PRESENTATION TO BE INCOMPREHENSIBLE, OR DO YOU HAVE SOME SORT OF RARE "POWERPOINT" DISABILITY?
© 2003 United Feature Syndicate, Inc.
8/9/03
ARE THERE ANY QUESTIONS ABOUT THE CONTENT?
THERE WAS CONTENT?